NORTH-HOLLAND

Backward Error for the Discrete-Time Algebraic Riccati Equation

Ji-guang Sun*
Department of Computing Science
Umeå University
S-901 87 Umeå, Sweden

Submitted by Volker Mehrmann

ABSTRACT

The normwise backward error of an approximate solution to the discrete-time algebraic Riccati equation is evaluated. The results are illustrated by using simple numerical examples. © Elsevier Science Inc., 1997

1. INTRODUCTION

For a linear system $Ax = b$, the normwise backward error of an approximate solution can be defined by [6, 7, 14]

$$\eta(\tilde{x}) = \min\left\{\epsilon : (A + \Delta A)\tilde{x} = b + \Delta b, \frac{\|\Delta A\|}{\alpha} \leqslant \varepsilon, \frac{\|\Delta b\|}{\beta} \leqslant \epsilon\right\},$$

* This work was supported by the Swedish Natural Science Research Council under contract M-AA/MA 06952-303 and the Department of Computing Science, Umeå University.

LINEAR ALGEBRA AND ITS APPLICATIONS 259:183–208 (1997)

655 Avenue of the Americas, New York, NY 10010

0024-3795/97/$17.00
PII S0024-3795(96)00283-2

where α and β are positive scalars, $\|\ \|$ denotes any consistent norm. If $\alpha = \|A\|$ and $\beta = \|b\|$, then $\eta(\tilde{x})$ is said to be the normwise relative backward error. Rigal and Gaches [14] derive the explicit expression

$$\eta(\tilde{x}) = \frac{\|\hat{r}\|}{\alpha\|\tilde{x}\| + \beta},$$

where $\hat{r} = b - A\tilde{x}$ is the associated residual. Moreover, Rigal and Gaches also determine the perturbations $\Delta A_{\min}$ and $\Delta b_{\min}$ where the optimal value $\eta(\tilde{x})$ is achieved.

From the definition of $\eta(\tilde{x})$ we see that the backward error $\eta(\tilde{x})$ of an approximate solution $\tilde{x}$ to $Ax = b$ is a measure of "smallest" perturbations $\Delta A/\alpha$ and $\Delta b/\beta$ such that $\tilde{x}$ is just the solution to the perturbed linear system $(A + \Delta A)\tilde{x} = b + \Delta b$. Rigal and Gaches's result covers an important aspect of the perturbation theory for linear systems.

In recent years, the study of backward errors of matrix equations has been developed considerably. Taking full account of the special structure of the Sylvester equation, Higham [6] evaluates the backward error of an approximate solution to the matrix equation, and determines the sensitivity of the equation to perturbations in the data. After that, Kågström [9] evaluates the normwise backward error of an approximate solution to the generalized Sylvester equation, and determines the sensitivity of the equation; Ghavimi and Laub [3] present a new backward error criterion, together with a sensitivity measure, for assessing solution accuracy of nonsymmetric and symmetric continuous-time algebraic Riccati equations.

The purpose of this paper is to evaluate the normwise backward error of an approximate solution to the discrete-time algebraic Riccati equation. The work was greatly influenced by the works of Higham in [6], and Ghavimi and Laub in [3].

Throughout this paper we use the following notation. The symbol $\mathscr{R}^{m\times n}$ denotes the set of real $m \times n$ matrices, and $\mathscr{R}^m = \mathscr{R}^{m\times 1}$. $\mathscr{S}^{n\times n}$ denotes the set of $n \times n$ symmetric matrices. A^T denotes the transpose of a matrix A, and $A^\dagger$ the Moore-Penrose inverse of A. I stands for the identity matrix, I_n for the identity matrix of order n, and 0 the null matrix. The positive definiteness (semidefiniteness) of a symmetric matrix A will be denoted by $A > 0$ ($A \geqslant 0$). $\|\ \|_F$ denotes the Frobenius norm, and $\|\ \|_2$ the Euclidean vector norm. For $A = (a_1, \ldots, a_n) = (\alpha_{ij}) \in \mathscr{C}^{n\times n}$ and a matrix B, $A \otimes B = (\alpha_{ij}B)$ is a Kronecker product, and vec A is a vector defined by vec $A = (a_1^T, \ldots, a_n^T)^T$. (See [4, Chapters 1 and 2] for properties of the Kronecker product and vec operation.)

Consider the discrete-time algebraic Riccati equation (DARE)

$$X - F^T XF + F^T XG_1\left(G_2 + G_1^T XG_1\right)^{-1} G_1^T XF - C^T C = 0, \quad (1.1)$$

where $F \in \mathscr{R}^{n\times n}$, $C \in \mathscr{R}^{r\times n}$, $G_1 \in \mathscr{R}^{n\times m}$, and $G_2^T = G_2 \in \mathscr{R}^{m\times m}$. In some control problems the matrix G_2 can be singular; but in this paper, as in [2, 5, 10, 11, 13], we assume $G_2 > 0$.

Let $G = G_1 G_2^{-1} G_1^T$, $H = C^T C$. Then the DARE (1.1) can be rewritten in the equivalent form

$$X - F^T X(I + GX)^{-1} F - H = 0, \quad (1.2)$$

where $G, H \geqslant 0$.

Throughout this paper we assume that (F, G_1) is a stabilizable pair (i.e., if $w^H G_1 = 0$ and $w^H F = \lambda w^H$ for some constant λ implies $|\lambda| < 1$ or $w = 0$, where w^H denotes the conjugate transpose of a complex vector w) and that (F, C) is a detectable pair (i.e., if (F^T, C^T) is stabilizable). It is known [1] that in such a case there exists a unique symmetric positive semidefinite (p.s.d.) solution X to the DARE (1.1), and the matrix $(I + GX)^{-1}F$ is stable, i.e., every eigenvalue λ_i of $(I + GX)^{-1}F$ satisfies $|\lambda_i| < 1$.

The DARE (1.1) arises in the solution of quadratic optimization and estimation problems in linear control theory. Several elegant numerical methods for solving the equation have been proposed (see, e,g., [2, 10, 11, 13]). Perturbation theory for the DARE (1.1) [or equivalently, for the DARE (1.2)] has also been studied by a certain number of authors [5, 8, 15, 16].

We assume that perturbations of G_1 and G_2 are transformed into perturbations of G and perturbations of C into perturbations of H. Thus, we consider the DARE (1.2) in this paper. Let $\tilde{X} \in \mathscr{S}^{n\times n}$ approximate the unique symmetric p.s.d. solution to the equation, and let ΔF, ΔG, and ΔH be the corresponding perturbations of the coefficient matrices F, G, and H in (1.2). A normwise backward error of the approximate solution $\tilde{X}$ can be defined by

$$\eta(\tilde{X}) = \min\left\{\epsilon : \frac{\|\Delta F\|_F}{\alpha} \leqslant \epsilon, \frac{\|\Delta G\|_F}{\beta} \leqslant \epsilon, \frac{\|\Delta H\|_F}{\gamma} \leqslant \epsilon\right\} \quad (1.3)$$

subject to

$$\Delta F \in \mathscr{R}^{n\times n}, \qquad \Delta G, \Delta H \in \mathscr{S}^{n\times n}, \quad (1.4)$$

$$\|\tilde{X}\|_2 \left\|(I + G\tilde{X})^{-1}\right\|_2 \|\Delta G\|_2 < 1, \quad (1.5)$$

and

$$\tilde{X} - (F + \Delta F)^T \tilde{X}\left[I + (G + \Delta G)\tilde{X}\right]^{-1}(F + \Delta F) - (H + \Delta H) = 0, \tag{1.6}$$

where α, β, and γ are positive scalars. If $\alpha = \|F\|_F$, $\beta = \|G\|_F$, and $\gamma = \|H\|_F$, then $\eta(\tilde{X})$ corresponds to a normwise relative backward error. Note that the constraint (1.5) guarantees the nonsingularity of the matrix $I + (G + \Delta G)\tilde{X}$.

From (1.3)–(1.6) we see that the backward error $\eta(\tilde{X})$ of an approximate solution $\tilde{X}$ to the DARE (1.2) is a measure of "smallest" perturbations $\Delta F/\alpha$, $\Delta G/\beta$, and $\Delta H/\gamma$ such that $\tilde{X}$ is just a symmetric solution to the perturbed DARE (1.6).

The difficult point for evaluating the backward error for the DARE (1.2) is that we are confronted with a nonlinear problem: the optimization problem (1.3) with the constraints (1.4)–(1.6), where the constraint (1.6) is nonlinear. Note that in the cases of Sylvester and continuous Riccati equations (studied by Higham [6] and by Ghavimi and Laub [3]), the associated constraints are linear.

In Section 2, we shall first transform the equation (1.6) to an equivalent form which is easy to handle. The main result of this paper is stated by Theorem 2.1, which presents lower and upper bounds for the backward error $\eta(\tilde{X})$. In Section 3 we present approximate bounds for $\eta(\tilde{X})$ by using the structured relative residual, and derive the structured condition number of the DARE (1.2). In Section 4 we provide two simple numerical examples to illustrate our results.

Let $A \in \mathscr{R}^{n\times n}$. Then we have [4, pp. 32–34]

$$\operatorname{vec} A^T = \Pi \operatorname{vec} A, \tag{1.7}$$

where the vec-permutation matrix Π is expressed by

$$\Pi = \sum_{k,l=1}^{n} e_k e_l^T \otimes e_l e_k^T, \tag{1.8}$$

in which e_k denotes the kth column of I_n.

2. ESTIMATES FOR $\eta(\tilde{X})$

The problem of deriving an explicit expression for the backward error $\eta(\tilde{X})$ defined by (1.3)–(1.6) is a difficult one. In this section we only give some estimates for $\eta(\tilde{X})$.

Let $\tilde{X} \in \mathscr{S}^{n\times n}$ be an approximate solution to the DARE (1.2). We now define another backward error $\eta^*(\tilde{X})$ of the approximate solution by

$$\eta^*(\tilde{X}) = \min \left\| \left(\frac{\Delta F}{\alpha}, \frac{\Delta G}{\beta}, \frac{\Delta H}{\gamma} \right) \right\|_F \quad \text{subject to (1.4)–(1.6).} \tag{2.1}$$

From the definitions (1.3)–(1.6) and (2.1),

$$\frac{1}{\sqrt{3}} \eta^*(\tilde{X}) \leqslant \eta(\tilde{X}) \leqslant \eta^*(\tilde{X}). \tag{2.2}$$

Consequently, the problem of estimating $\eta(\tilde{X})$ can be reduced to estimate $\eta^*(\tilde{X})$.

2.1. *An Equivalent Form of the Equation (1.6)*

Assume that the matrix $I + G\tilde{X}$ is nonsingular. Then the equation (1.6) can be written as

$$\tilde{X} - (F + \Delta F)^T \tilde{X} \left[I + (I + G\tilde{X})^{-1} \Delta G \tilde{X} \right]^{-1} (I + G\tilde{X})^{-1} (F + \Delta F) - (H + \Delta H) = 0. \tag{2.3}$$

In this subsection we transform the equation (2.3) to an equivalent form which is easy to handle. In the course of the transformation, the matrix relations

$$(I + A)^{-1} = I - A(I + A)^{-1}, \qquad B(I + AB)^{-1} = (I + BA)^{-1} B \tag{2.4}$$

are used again and again.

Define the matrices $\tilde{L}$ and $\tilde{K}$ by

$$\tilde{L} = \tilde{X}(I + G\tilde{X})^{-1} \in \mathscr{S}^{n\times n}, \qquad \tilde{K} = \tilde{L}F \in \mathscr{R}^{n\times n}. \tag{2.5}$$

Since

$$\begin{aligned}
&\tilde{X}\Big[I + (I + G\tilde{X})^{-1}\Delta G\tilde{X}\Big]^{-1}(I + G\tilde{X})^{-1} \\
&\quad = \tilde{X}\Big\{I - (I + G\tilde{X})^{-1}\Delta G\Big[I + \tilde{X}(I + G\tilde{X})^{-1}\Delta G\Big]^{-1}\tilde{X}\Big\}(I + G\tilde{X})^{-1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad [\text{by (2.4)}] \\
&\quad = \Big[I - \tilde{L}\Delta G(I + \tilde{L}\Delta G)^{-1}\Big]\tilde{L} = \tilde{L} - p(\Delta G),
\end{aligned}$$

where $p(\Delta G)$ is defined by

$$p(\Delta G) = \tilde{L}\Delta G(I + \tilde{L}\Delta G)^{-1}\tilde{L} \in \mathscr{S}^{n\times n}, \tag{2.6}$$

we can rewrite (2.3) as

$$\tilde{X} - (F + \Delta F)^T\big[\tilde{L} - p(\Delta G)\big](F + \Delta F) - (H + \Delta H) = 0,$$

or equivalently,

$$\begin{aligned}
&\tilde{X} - H - \Delta H - F^T\tilde{L}F - F^T\tilde{L}\Delta F + F^Tp(\Delta G)F + F^Tp(\Delta G)\Delta F \\
&\quad - \Delta F^T\tilde{L}F - \Delta F^T\tilde{L}\Delta F + \Delta F^Tp(\Delta G)F + \Delta F^Tp(\Delta G)\Delta F = 0.
\end{aligned} \tag{2.7}$$

Observe the following facts: By (2.5) and (2.6) we have

$$\begin{aligned}
F^T\tilde{L}\Delta F &= \tilde{K}^T\Delta F, \qquad \Delta F^T\tilde{L}F = \Delta F^T\tilde{K}, \\
F^Tp(\Delta G)F &= \tilde{K}^T\Delta G(I + \tilde{L}\Delta G)^{-1}\tilde{K} \\
&= \tilde{K}^T\Delta G\Big[I - \tilde{L}\Delta G(I + \tilde{L}\Delta G)^{-1}\Big]\tilde{K} \\
&= \tilde{K}^T\Delta G\tilde{K} - \tilde{K}^T\Delta G\tilde{L}\Delta G(I + \tilde{L}\Delta G)^{-1}\tilde{K}, \\
F^Tp(\Delta G)\Delta F &= F^T\tilde{L}\Delta G(I + \tilde{L}\Delta G)^{-1}\tilde{L}\Delta F \\
&= \tilde{K}^T\Delta G(I + \tilde{L}\Delta G)^{-1}\tilde{L}\Delta F, \\
\Delta F^Tp(\Delta G)F &= \Delta F^T\tilde{L}\Delta G(I + \tilde{L}\Delta G)^{-1}\tilde{L}F \\
&= \Delta F^T\tilde{L}\Delta G(I + \tilde{L}\Delta G)^{-1}\tilde{K},
\end{aligned}$$

and

$$\begin{aligned}
&\Delta F^T p(\Delta G)\,\Delta F - \Delta F^T \tilde{L}\,\Delta F \\
&\quad = \Delta F^T \tilde{L}\,\Delta G\,(I + \tilde{L}\,\Delta G)^{-1}\tilde{L}\,\Delta F - \Delta F^T \tilde{X}(I + G\tilde{X})^{-1}\Delta F \\
&\quad = -\Delta F^T (I + \tilde{L}\,\Delta G)^{-1}\tilde{L}\,\Delta F.
\end{aligned}$$

Hence, the equation (2.7) is equivalent to

$$\tilde{K}^T \Delta F + \Delta F^T \tilde{K} - \tilde{K}^T \Delta G \tilde{K} + \Delta H = \hat{R} + q(\Delta F, \Delta G), \tag{2.8}$$

where $\hat{R}$, the residual of the DARE (1.2) with respect to $\tilde{X}$, is defined by

$$\hat{R} = \tilde{X} - F^T \tilde{L} F - H = \tilde{X} - F^T \tilde{X}(I + G\tilde{X})^{-1} F - H, \tag{2.9}$$

and $q(\Delta F, \Delta G)$ is defined by

$$\begin{aligned}
&q(\Delta F, \Delta G) \\
&\quad = -\tilde{K}^T \Delta G \tilde{L}\,\Delta G\,(I + \tilde{L}\,\Delta G)^{-1}\tilde{K} + \tilde{K}^T \Delta G\,(I + \tilde{L}\,\Delta G)^{-1}\tilde{L}\,\Delta F \\
&\qquad + \Delta F^T \tilde{L}\,\Delta G\,(I + \tilde{L}\,\Delta G)^{-1}\tilde{K} - \Delta F^T (I + \tilde{L}\,\Delta G)^{-1}\tilde{L}\,\Delta F,
\end{aligned} \tag{2.10}$$

in which $\tilde{L}$ and $\tilde{K}$ are defined by (2.5). Note that $q(\Delta F, \Delta G) \in \mathscr{S}^{n\times n}$.

Further, by using a technique described by Higham [6], we transform (2.8) to a simpler form. Let $\tilde{K}$ be the matrix defined by (2.5), and let

$$\tilde{K} = U\Sigma V^T \quad \text{with} \quad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n), \quad \sigma_1 \geqslant \cdots \geqslant \sigma_n \geqslant 0 \tag{2.11}$$

be a singular-value decomposition of $\tilde{K}$, where $U, V \in \mathscr{R}^{n\times n}$ are orthogonal. Substituting (2.11) into (2.8), and setting

$$\begin{aligned}
&\delta F = U^T \Delta F V, \qquad \delta G = U^T \Delta G\, U, \qquad \delta H = V^T \Delta H V, \\
&\tilde{R} = V^T \hat{R} V, \qquad \tilde{q}(\Delta F, \Delta G) = V^T q(\Delta F, \Delta G) V,
\end{aligned} \tag{2.12}$$

then (2.8) reduces to

$$\Sigma\,\delta F + \delta F^T\,\Sigma - \Sigma\,\delta G\,\Sigma + \delta H = \tilde{R} + \tilde{q}(\Delta F, \Delta G). \tag{2.13}$$

By (1.7) and (1.8), the equation (2.13) is equivalent to the nonlinear system

$$T\begin{pmatrix} \dfrac{\operatorname{vec}\delta F}{\alpha} \\ \dfrac{\operatorname{vec}\delta G}{\beta} \\ \dfrac{\operatorname{vec}\delta H}{\gamma} \end{pmatrix} = \operatorname{vec}\tilde{R} + \operatorname{vec}\tilde{q}(\Delta F, \Delta G), \tag{2.14}$$

where

$$T = (T_1, T_2, T_3) \in \mathscr{R}^{n^2\times 3n^2} \tag{2.15}$$

with

$$T_1 = \alpha\big[I_n \otimes \Sigma + (\Sigma \otimes I_n)\Pi\big], \qquad T_2 = -\beta\Sigma \otimes \Sigma, \qquad T_3 = \gamma I_{n^2}, \tag{2.16}$$

in which Π is the vec-permutation matrix expressed by (1.8).

2.2. An Upper Bound for $\eta^*(\tilde{X})$

Consider the linear system

$$\begin{pmatrix} \dfrac{\operatorname{vec}\delta F}{\alpha} \\ \dfrac{\operatorname{vec}\delta G}{\beta} \\ \dfrac{\operatorname{vec}\delta H}{\gamma} \end{pmatrix} = T^{\dagger}\big[\operatorname{vec}\tilde{R} + \operatorname{vec}\tilde{q}(\Delta F, \Delta G)\big], \tag{2.17}$$

where δF, δG, δH, $\tilde{R}$, and $\tilde{q}(\Delta F, \Delta G)$ are defined by (2.12), in which $\Delta F \in \mathscr{R}^{n\times n}$, and ΔG, $\Delta H \in \mathscr{S}^{n\times n}$. Since the $n^2 \times 3n^2$ matrix T is full rank

(because it is assumed that $\gamma > 0$), we have $TT^{\dagger} = I_{n^2}$, so multiplying the equation (2.17) on the left by T yields the equation (2.14). This shows that any solution to the equation (2.17) is a solution to the equation (2.14). For this reason, if

$$\left(\frac{(\operatorname{vec} \delta F_*)^T}{\alpha}, \frac{(\operatorname{vec} \delta G_*)^T}{\beta}, \frac{(\operatorname{vec} \delta H_*)^T}{\gamma} \right)^T$$

is a solution to (2.17), then

$$\eta^*(\tilde{X}) \leqslant \left\| \begin{pmatrix} \dfrac{\operatorname{vec} \delta F_*}{\alpha} \\ \dfrac{\operatorname{vec} \delta G_*}{\beta} \\ \dfrac{\operatorname{vec} \delta H_*}{\gamma} \end{pmatrix} \right\|_2 = \left\| \left(\frac{\Delta F_*}{\alpha}, \frac{\Delta G_*}{\beta}, \frac{\Delta H_*}{\gamma} \right) \right\|_F, \quad (2.18)$$

where $\eta^*(\tilde{X})$ is defined by (2.1), and

$$\Delta F_* = U \delta F_* V^T, \qquad \Delta G_* = U \delta G_* U^T, \qquad \Delta H_* = V \delta H_* V^T.$$

It is evident that the nonlinear system (2.17) can be regarded as a continuous mapping $\mathscr{M} : \mathscr{R}^{n\times n} \oplus \mathscr{S}^{n\times n} \oplus \mathscr{S}^{n\times n} \to \mathscr{R}^{n\times n} \oplus \mathscr{S}^{n\times n} \oplus \mathscr{S}^{n\times n}$. Note that the set $\mathscr{R}^{n\times n} \oplus \mathscr{S}^{n\times n} \oplus \mathscr{S}^{n\times n}$ with the norm $\| \ \|_F$ is a Banach space. From (2.17) we see that the mapping $\mathscr{M}$ satisfies

$$\left\| \left(\frac{\Delta F}{\alpha}, \frac{\Delta G}{\beta}, \frac{\Delta H}{\gamma} \right) \right\|_F \leqslant \rho + \frac{1}{\tau} \| q(\Delta F, \Delta G) \|_F, \quad (2.19)$$

where ρ and τ are defined by

$$\rho = \| T^{\dagger} \operatorname{vec} \tilde{R} \|_2, \qquad \tau = \| T^{\dagger} \|_2^{-1}. \quad (2.20)$$

By (2.10), we have

$$\begin{aligned}\|q(\Delta F, \Delta G)\|_F &\leqslant \frac{\|\tilde{L}\|_2}{1-\beta\|\tilde{L}\|_2\|\Delta G/\beta\|_F}\left(\alpha\left\|\frac{\Delta F}{\alpha}\right\|_F+\beta\|\tilde{K}\|_2\left\|\frac{\Delta G}{\beta}\right\|_F\right)^2\\ &\leqslant \frac{(\alpha^2+\beta^2\|\tilde{K}\|_2^2)\|\tilde{L}\|_2}{1-\beta\|\tilde{L}\|_2\|\Delta G/\beta\|_F}\left(\left\|\frac{\Delta F}{\alpha}\right\|_F^2+\left\|\frac{\Delta G}{\beta}\right\|_F^2\right)\\ &\leqslant \frac{\mu\left\|\left(\frac{\Delta F}{\alpha},\frac{\Delta G}{\beta},\frac{\Delta H}{\gamma}\right)\right\|_F^2}{1-\nu\left\|\left(\frac{\Delta F}{\alpha},\frac{\Delta G}{\beta},\frac{\Delta H}{\gamma}\right)\right\|_F},\end{aligned} \tag{2.21}$$

where μ and ν are defined by

$$\mu=(\alpha^2+\beta^2\|\tilde{K}\|_2^2)\|\tilde{L}\|_2, \qquad \nu=\beta\|\tilde{X}\|_2\|(I+G\tilde{X})^{-1}\|_2, \tag{2.22}$$

and it is assumed that

$$1-\nu\left\|\left(\frac{\Delta F}{\alpha},\frac{\Delta G}{\beta},\frac{\Delta H}{\gamma}\right)\right\|_F>0. \tag{2.23}$$

Obviously, the condition (2.23) implies the constraint (1.5). Combining (2.19) with (2.21) shows that the mapping $\mathscr{M}$ satisfies

$$\left\|\left(\frac{\Delta F}{\alpha},\frac{\Delta G}{\beta},\frac{\Delta H}{\gamma}\right)\right\|_F\leqslant\rho+\frac{\mu\left\|\left(\frac{\Delta F}{\alpha},\frac{\Delta G}{\beta},\frac{\Delta H}{\gamma}\right)\right\|_F^2}{\tau\left(1-\nu\left\|\left(\frac{\Delta F}{\alpha},\frac{\Delta G}{\beta},\frac{\Delta H}{\gamma}\right)\right\|_F\right)}, \tag{2.24}$$

or equivalently,

$$\begin{aligned}(\tau\nu+\mu)&\left\|\left(\frac{\Delta F}{\alpha},\frac{\Delta G}{\beta},\frac{\Delta H}{\gamma}\right)\right\|_F^2-\tau(1+\nu\rho)\left\|\left(\frac{\Delta F}{\alpha},\frac{\Delta G}{\beta},\frac{\Delta H}{\gamma}\right)\right\|_F\\ &+\tau\rho\geqslant 0,\end{aligned} \tag{2.25}$$

where ρ, τ are defined by (2.20), and μ, ν by (2.22).

Consider the equation

$$(\tau\nu + \mu)\xi^2 - \tau(1 + \nu\rho)\xi + \tau\rho = 0. \tag{2.26}$$

It is easy to verify that if ρ satisfies

$$\rho \leqslant \frac{\tau}{\tau\nu + 2\mu + \sqrt{(\tau\nu + 2\mu)^2 - \tau^2\nu^2}}, \tag{2.27}$$

then

$$\tau^2(1 + \nu\rho)^2 - 4\tau(\tau\nu + \mu)\rho \geqslant 0, \tag{2.28}$$

and in this case the positive scalar ξ_* expressed by

$$\xi_* = \frac{2\tau\rho}{\tau(1 + \nu\rho) + \sqrt{\tau^2(1 + \nu\rho)^2 - 4\tau(\tau\nu + \mu)\rho}} \equiv u(\rho) \tag{2.29}$$

is a solution to (2.26).

Let

$$\mathscr{S}_{\xi_*} = \left\{ \left(\frac{\Delta F}{\alpha}, \frac{\Delta G}{\beta}, \frac{\Delta H}{\gamma} \right) \in \mathscr{R}^{n\times n} \oplus \mathscr{S}^{n\times n} \oplus \mathscr{S}^{n\times n} : \right.$$

$$\left. \left\| \left(\frac{\Delta F}{\alpha}, \frac{\Delta G}{\beta}, \frac{\Delta H}{\gamma} \right) \right\|_F \leqslant \xi_* \right\}. \tag{2.30}$$

Obviously, $\mathscr{S}_{\xi_*}$ is a bounded closed convex set of $\mathscr{R}^{n\times n} \oplus \mathscr{S}^{n\times n} \oplus \mathscr{S}^{n\times n}$, and the relation (2.24) shows that the continuous mapping $\mathscr{M}$ maps $\mathscr{S}_{\xi_*}$ into $\mathscr{S}_{\xi_*}$. By the Schauder fixed-point theorem (see, e.g., [12, §6.3]), the mapping $\mathscr{M}$ has a fixed point in $\mathscr{S}_{\xi_*}$, i.e., $\mathscr{M}$ has a fixed point $(\Delta F_*/\alpha, \Delta G_*/\beta, \Delta H_*/\gamma)$ satisfying

$$\left\| \left(\frac{\Delta F_*}{\alpha}, \frac{\Delta G_*}{\beta}, \frac{\Delta H_*}{\gamma} \right) \right\|_F \leqslant \xi_*, \tag{2.31}$$

where ξ_* is expressed by (2.29).

Thus, we have proved that under the conditions (2.23) and (2.27) there is a solution $((\text{vec}\,\delta F_*)^T/\alpha, (\text{vec}\,\delta G_*)^T/\beta, (\text{vec}\,\delta H_*)^T/\gamma)^T$ to the equation (2.17) such that

$$\left\| \begin{pmatrix} \dfrac{\text{vec}\,\delta F_*}{\alpha} \\ \dfrac{\text{vec}\,\delta G_*}{\beta} \\ \dfrac{\text{vec}\,\delta H_*}{\gamma} \end{pmatrix} \right\|_2 \leqslant \xi_*. \tag{2.32}$$

Observe that any solution $((\text{vec}\,\delta F)^T/\alpha, (\text{vec}\,\delta G)^T/\beta, (\text{vec}\,\delta H)^T/\gamma)^T$ to the equation (2.17) [where $\delta F \in \mathscr{R}^{n\times n}$ and $\delta G, \delta H \in \mathscr{S}^{n\times n}$, i.e., $\delta F, \delta G, \delta H$ satisfy (1.4)] is a solution to the equation (2.14), and the equation (2.14) is an equivalent form of the equation (1.6); moreover, the condition (2.23) implies the condition (1.5). Hence, the solution $((\text{vec}\,\delta F_*)^T/\alpha, (\text{vec}\,\delta G_*)^T/\beta, (\text{vec}\,\delta H_*)^T/\gamma)^T$ satisfies (1.4)–(1.6). Combining (2.32) with (2.18) gives

$$\eta^*(\tilde{X}) \leqslant \xi_* = u(\rho). \tag{2.33}$$

Note that if ρ satisfies

$$\rho < 1/\nu, \tag{2.34}$$

then any solution $((\text{vec}\,\delta F)^T/\alpha, (\text{vec}\,\delta G)^T/\beta, (\text{vec}\,\delta H)^T/\gamma)^T$ of (2.17) in $\mathscr{S}_{\xi_*}$ satisfies (2.23). In fact, by (2.29) and (2.30), we have

$$\nu \left\| \left(\frac{\Delta F}{\alpha}, \frac{\Delta G}{\beta}, \frac{\Delta H}{\gamma} \right) \right\|_F \leqslant \nu\xi_* \leqslant \frac{2\nu\rho}{1+\nu\rho} < 1 \qquad [\text{by (2.34)}]. \tag{2.35}$$

Consequently, the condition (2.23) can be replaced by (2.34). Moreover, the conditions (2.27) and (2.34) can be expressed by

$$\rho < \min\left\{ \frac{1}{\nu}, \frac{\tau}{\tau\nu + 2\mu + \sqrt{(\tau\nu+2\mu)^2 - \tau^2\nu^2}} \right\}. \tag{2.36}$$

2.3. *A Lower Bound for* $\eta^*(\tilde{X})$

Suppose that the matrices $\Delta F_{\min}$, $\Delta G_{\min}$, and $\Delta H_{\min}$ satisfy the constraints (1.4)–(1.6), and

$$\eta^*(\tilde{X}) = \left\| \left(\frac{\Delta F_{\min}}{\alpha}, \frac{\Delta G_{\min}}{\beta}, \frac{\Delta H_{\min}}{\gamma} \right) \right\|_F. \tag{2.37}$$

By the result of Section 2.2, under the condition (2.36) we have

$$\left\| \begin{pmatrix} \dfrac{\operatorname{vec} \delta F_{\min}}{\alpha} \\ \dfrac{\operatorname{vec} \delta G_{\min}}{\beta} \\ \dfrac{\operatorname{vec} \delta H_{\min}}{\gamma} \end{pmatrix} \right\|_2 = \left\| \left(\frac{\Delta F_{\min}}{\alpha}, \frac{\Delta G_{\min}}{\beta}, \frac{\Delta H_{\min}}{\gamma} \right) \right\|_F \leqslant \xi_*, \tag{2.38}$$

where $\xi_* = u(\rho)$ is expressed by (2.29), and where $\delta F_{\min} \in \mathscr{R}^{n\times n}$ and $\delta G_{\min}, \delta H_{\min} \in \mathscr{S}^{n\times n}$ are defined by

$$\delta F_{\min} = U^T \Delta F_{\min} V, \qquad \delta G_{\min} = U^T \Delta G_{\min} U, \qquad \delta H_{\min} = V^T \Delta H_{\min} V.$$

Since (2.14) is an equivalent form of (1.5), by the assumption the matrices $\delta F_{\min}$, $\delta G_{\min}$, $\delta H_{\min}$ satisfy

$$T \begin{pmatrix} \dfrac{\operatorname{vec} \delta F_{\min}}{\alpha} \\ \dfrac{\operatorname{vec} \delta G_{\min}}{\beta} \\ \dfrac{\operatorname{vec} \delta H_{\min}}{\gamma} \end{pmatrix} = \operatorname{vec} \tilde{R} + \operatorname{vec} \tilde{q}(\Delta F_{\min}, \Delta G_{\min}). \tag{2.39}$$

Let

$$T = W(\Omega, 0) Z^T \qquad \text{with} \quad \Omega = \operatorname{diag}(\omega_1, \ldots, \omega_{n^2}), \quad \omega_1 \geqslant \cdots \geqslant \omega_{n^2} > 0, \tag{2.40}$$

be a singular-value decomposition of T, where $W \in \mathscr{R}^{n^2 \times n^2}$ and $Z \in \mathscr{R}^{3n^2 \times 3n^2}$ are orthogonal. Substituting (2.40) into (2.39), and letting

$$Z^T \begin{pmatrix} \dfrac{\operatorname{vec} \delta F_{\min}}{\alpha} \\ \dfrac{\operatorname{vec} \delta G_{\min}}{\beta} \\ \dfrac{\operatorname{vec} \delta H_{\min}}{\gamma} \end{pmatrix} = \begin{pmatrix} v \\ * \\ * \end{pmatrix} \quad \text{with} \quad v \in \mathscr{R}^{n^2}, \tag{2.41}$$

we get

$$v = \Omega^{-1} W^T \operatorname{vec} \tilde{R} + \Omega^{-1} W^T \operatorname{vec} \tilde{q}(\Delta F_{\min}, \Delta G_{\min}). \tag{2.42}$$

Combining (2.40)–(2.42) with (2.37) and (2.38) gives

$$\begin{aligned} \eta^*(\tilde{X}) &= \left\| \begin{pmatrix} \dfrac{\operatorname{vec} \delta F_{\min}}{\alpha} \\ \dfrac{\operatorname{vec} \delta G_{\min}}{\beta} \\ \dfrac{\operatorname{vec} \delta H_{\min}}{\gamma} \end{pmatrix} \right\|_2 \geqslant \|v\|_2 \\ &\geqslant \|\Omega^{-1} W^T \operatorname{vec} \tilde{R}\|_2 - \left\| \Omega^{-1} W^T \operatorname{vec} \tilde{q}(\Delta F_{\min}, \Delta G_{\min}) \right\|_2 \\ &\geqslant \|T^\dagger \operatorname{vec} \tilde{R}\|_2 - \|T^\dagger\|_2 \left\| \tilde{q}(\Delta F_{\min}, \Delta G_{\min}) \right\|_F. \end{aligned} \tag{2.43}$$

The relations (2.43), (2.20), (2.21), and (2.38) show that under the conditions (2.36) we have

$$\eta^*(\tilde{X}) \geqslant \rho - \frac{\mu \xi_*^2}{\tau(1 - \nu \xi_*)} = \rho - \frac{\mu [u(\rho)]^2}{\tau[1 - \nu u(\rho)]} \equiv l(\rho), \tag{2.44}$$

where $\xi_* = u(\rho)$ is expressed by (2.29).

Note that the scalar $l(\rho)$ defined by (2.44) satisfies $l(\rho) \geqslant 0$. Moreover, if $\tilde{X} \neq 0$ and $\rho > 0$, then $l(\rho) > 0$. These facts can be proved as follows. By

(2.35), $1 - \nu\xi_* > 0$. Consequently, it only needs to show the inequality

$$\rho \geqslant \frac{\mu\xi_*^2}{\tau(1 - \nu\xi_*)}, \qquad \text{i.e.,} \quad \mu\xi_*^2 + \tau\nu\rho\xi_* - \tau\rho \leqslant 0,$$

or equivalently, to show the inequality

$$\xi_* \leqslant \frac{2\tau\rho}{\tau\nu\rho + \sqrt{(\tau\nu\rho)^2 + 4\tau\mu\rho}}. \tag{2.45}$$

Observe the following facts: (i) by (2.29) we have

$$\xi_* \leqslant \frac{2\rho}{1 + \nu\rho}; \tag{2.46}$$

(ii) it is easy to verify that the inequality

$$\frac{2\rho}{1 + \nu\rho} \leqslant \frac{2\tau\rho}{\tau\nu\rho + \sqrt{(\tau\nu\rho)^2 + 4\tau\mu\rho}} \tag{2.47}$$

is equivalent to

$$\tau^2 - (\tau\nu\rho)^2 - 4\tau\mu\rho \geqslant 0, \tag{2.48}$$

and the inequality (2.48) holds because we have

$$\begin{aligned} \tau^2 - (\tau\nu\rho)^2 - 4\tau\mu\rho &\geqslant 2\tau^2\nu\rho(1 - \nu\rho) \qquad [\text{by (2.28)}] \\ &\geqslant 0 \qquad [\text{by (2.35)}]. \end{aligned}$$

Hence, combining (2.47) with (2.46) shows (2.45). Moreover, if $\bar{X} \neq 0$ and $\rho > 0$, then the last inequality becomes

$$2\tau^2\nu\rho(1 - \nu\rho) > 0 \qquad [\text{by (2.35) and } \nu > 0,\ \rho > 0].$$

Therefore, in this case we have $l(\rho) > 0$.

2.4. Lower and Upper Bounds for $\eta(\tilde{X})$

We now summarize our result of this section as follows:

THEOREM 2.1. *Let $\tilde{X} \in \mathcal{S}^{n\times n}$ approximate the unique symmetric p.s.d. solution to the DARE* (1.2), *and let $\eta(\tilde{X})$ be the normwise backward error of $\tilde{X}$ defined by* (1.3)–(1.6). *Assume that the matrix $I + G\tilde{X}$ is nonsingular. Moreover, define the matrices $\tilde{K}$, $\tilde{L}$, μ, and ν by*

$$\tilde{K} = \tilde{X}(I + G\tilde{X})^{-1}F, \qquad \tilde{L} = \tilde{X}(I + G\tilde{X})^{-1},$$

$$\mu = (\alpha^2 + \beta^2\|\tilde{K}\|_2^2)\|\tilde{L}\|_2, \qquad \nu = \beta\|\tilde{X}\|_2\|(I + G\tilde{X})^{-1}\|_2.$$

By using the singular-value decomposition $\tilde{K} = U\Sigma V^T$ of $\tilde{K}$, define an $n^2 \times 3n^2$ matrix T by (2.15)–(2.16), *and define τ, ρ by*

$$\tau = \|T^\dagger\|_2^{-1}, \qquad \rho = \|T^\dagger \operatorname{vec} \tilde{R}\|_2, \tag{2.49}$$

where $\tilde{R}$ is defined by

$$\tilde{R} = V^T\hat{R}V \qquad \text{with} \quad \hat{R} = \tilde{X} - F^T\tilde{X}(I + G\tilde{X})^{-1}F - H.$$

If ρ satisfies

$$\rho < \min\left\{\frac{1}{\nu}, \frac{\tau}{\tau\nu + 2\mu + \sqrt{(\tau\nu + 2\mu)^2 - \tau^2\nu^2}}\right\}, \tag{2.50}$$

then

$$\frac{1}{\sqrt{3}}l(\rho) \leqslant \eta(\tilde{X}) \leqslant u(\rho), \tag{2.51}$$

where

$$u(\rho) = \frac{2\tau\rho}{\tau(1 + \nu\rho) + \sqrt{\tau^2(1 + \nu\rho)^2 - 4\tau(\tau\nu + \mu)\rho}}, \tag{2.52}$$

and

$$l(\rho) = \rho - \frac{\mu[u(\rho)]^2}{\tau[1 - \nu u(\rho)]}. \tag{2.53}$$

The function $u(\rho)$ expressed by (2.52) has the Taylor expansion

$$u(\rho) = \rho + \frac{\mu}{\tau}\rho^2 + \left[\nu \cdot \frac{\mu}{\tau} + 2\left(\frac{\mu}{\tau}\right)^2\right]\rho^3 + O(\rho^4), \qquad \rho \to 0,$$

and the function $l(\rho)$ expressed by (2.53) has the Taylor expansion

$$l(\rho) = \rho - \frac{\mu}{\tau}\rho^2 - \left[\nu \cdot \frac{\mu}{\tau} + 2\left(\frac{\mu}{\tau}\right)^2\right]\rho^3 + O(\rho^4), \qquad \rho \to 0.$$

Consequently, for sufficiently small ρ we have the approximate estimates

$$\frac{1}{\sqrt{3}}\rho \lesssim \eta(\tilde{X}) \lesssim \rho. \tag{2.54}$$

3. TWO REMARKS

3.1. *Backward Error and Structured Relative Residual*

Taking full account of the special structure of the continuous-time algebraic Riccati equation, Ghavimi and Laub [3] have defined a new expression for the relative residual (they refer to the new expression as the structured relative residual for the equation), and proved that "backward error = relative residual" holds not only for linear system, but also in the case of continuous-time Riccati equations. In this subsection we show that the idea and conclusion of [3] are also suitable to DAREs.

Let $\tilde{X} \in \mathscr{S}^{n\times n}$ approximate the unique symmetric p.s.d. solution to the DARE (1.2). By Section 2.1, the perturbation equation (1.6) can be written as (2.13), where Σ, δF, δG, δH, $\tilde{R}$, and $\tilde{q}(\Delta F, \Delta G)$ are defined by (2.11)–(2.12). On dropping higher-order terms, the equation (2.13) becomes

$$\Sigma\,\delta F + \delta F^T\,\Sigma - \Sigma\,\delta G\,\Sigma + \delta H = \tilde{R}, \tag{3.1}$$

or equivalently, for $i, j = 1, \ldots, n$ we have

$$\sigma_i(U^T\Delta F V)_{ij} + \sigma_j\big((U^T\Delta F V)^T\big)_{ij} - \sigma_i\sigma_j(U^T\Delta G U)_{ij} + (V^T\Delta H V)_{ij}$$

$$= (\tilde{R})_{ij} \equiv r_{ij}.$$

By [3], define ρ_{ij} by

$$\rho_{ij} = \alpha(\sigma_i + \sigma_j) + \beta\sigma_i\sigma_j + \gamma \qquad \forall i, j, \tag{3.2}$$

and define the structured relative residual η_R by

$$\eta_R = \sqrt{\sum_{i,j=1}^{n}\left(\frac{r_{ij}}{\rho_{ij}}\right)^2}. \tag{3.3}$$

Consistently imitating the proof given by [3, Section 5], we can from (3.1)–(3.3) get approximate bounds for $\eta(\tilde{X})$ in terms of the associated structured relative residual η_R:

$$\eta_R \lesssim \eta(\tilde{X}) \lesssim 2.12\eta_R. \tag{3.4}$$

3.2. Structured Condition Number

In this subsection we use the technique described by Gudmundsson et al. [5] and Higham [6] to derive the structured condition number of the DARE (1.2).

Suppose that $\Delta F \in \mathscr{R}^{n\times n}$ and $\Delta G, \Delta H \in \mathscr{S}^{n\times n}$ are sufficiently small perturbations of F, G, H, respectively. Consider the perturbed DARE

$$X + \Delta X - (F + \Delta F)^T(X + \Delta X)[I + (G + \Delta G)(X + \Delta X)]^{-1}(F + \Delta F)$$

$$-(H + \Delta H) = 0. \tag{3.5}$$

By expanding and dropping the higher-order terms, the equation (3.5) becomes

$$\Delta X - \left[(I + GX)^{-1}F\right]^T \Delta X \left[(I + GX)^{-1}F\right]$$
$$= \Delta F^T \left[X(I + GX)^{-1}F\right] + \left[X(I + GX)^{-1}F\right]^T \Delta F$$
$$- \left[X(I + GX)^{-1}F\right]^T \Delta G \left[X(I + GX)^{-1}F\right] + \Delta H. \quad (3.6)$$

Define the matrices K and Φ by

$$K = X(I + GX)^{-1}F, \qquad \Phi = (I + GX)^{-1}F. \quad (3.7)$$

Then the equation (3.6) can be written in the form of the linear system

$$P \operatorname{vec} \Delta X = Q \begin{pmatrix} \dfrac{\operatorname{vec} \Delta F}{\alpha} \\ \dfrac{\operatorname{vec} \Delta G}{\beta} \\ \dfrac{\operatorname{vec} \Delta H}{\gamma} \end{pmatrix},$$

where

$$P = I_{n^2} - \Phi^T \otimes \Phi^T, \qquad Q = (Q_1, Q_2, Q_3) \in \mathscr{C}^{n^2 \times 3n^2} \quad (3.8)$$

with

$$Q_1 = \alpha\left[(K^T \otimes I_n)\Pi + I_n \otimes K^T\right],$$
$$Q_2 = -\beta K^T \otimes K^T, \qquad Q_3 = \gamma I_{n^2}, \qquad \alpha, \beta, \gamma > 0, \quad (3.9)$$

in which Π is the vec-permutation matrix expressed by (1.8). If we measure the perturbations in the data by

$$\epsilon = \max\left\{\frac{\|\Delta F\|_F}{\alpha}, \frac{\|\Delta G\|_F}{\beta}, \frac{\|\Delta H\|_F}{\gamma}\right\},$$

then we have

$$\frac{\|\Delta X\|_F}{\|X\|_F} \leqslant \sqrt{3}\,\psi\epsilon \qquad \text{(to first order in } \epsilon\text{)}, \tag{3.10}$$

where the scalar ψ defined by

$$\psi = \|P^{-1}Q\|_2/\|X\|_F \tag{3.11}$$

can be called the structured condition number of the DARE (1.2).

In particular, if we take $\alpha = \|F\|_F$, $\beta = \|G\|_F$, and $\gamma = \|H\|_F$ in (3.9), then the corresponding condition number ψ expressed by (3.11) can be called the relative condition number of the DARE (1.2). The relative condition number of the DARE has been derived by Gudmundsson et al. [5].

4. NUMERICAL EXAMPLES

In Sections 2 and 3 we have derived several estimates for the backward error $\eta(\tilde{X})$ of approximate solutions $\tilde{X}$ to the DARE (1.2), such as [see (2.51), (2.54), and (3.4)]:

$$\frac{l(\rho)}{\sqrt{3}} \leqslant \eta(\tilde{X}) \leqslant u(\rho), \tag{4.1}$$

and

$$\rho/\sqrt{3} \lesssim \eta(\tilde{X}) \lesssim \rho, \qquad \eta_R \lesssim \eta(\tilde{X}) \lesssim 2.12\eta_R, \tag{4.2}$$

where ρ, $u(\rho)$, $l(\rho)$, and η_R are defined by (2.49), (2.52), (2.53), and (3.3), respectively. The estimates (4.1) are nonlinear estimates, and the estimates (4.2) are linear estimates.

In this section, we use two simple numerical examples to illustrate the estimates. All computations were performed using MATLAB, version 4.2c, implemented on a SALT. The relative machine precision reported by MATLAB is 2.2204×10^{-16}.

EXAMPLE 4.1 [16, Example 6.2]. Consider the DARE (1.2) with the coefficient matrices

$$F = \begin{pmatrix} 0 & 10^m \\ 0 & 0 \end{pmatrix}, \qquad H = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = C^TC \quad \text{with } C = (0,1),$$

and

$$G = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = G_1 G_2^{-1} G_1^T \quad \text{with} \quad G_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } G_2 = 1.$$

It can be checked that the pair (F, G_1) is stabilizable and the pair (F, C) is detectable. Moreover, for all m, the symmetric p.s.d. solution to the DARE (1.2) is

$$X = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

and the matrix

$$(I + GX)^{-1} F = \begin{pmatrix} 0 & 10^m \\ 0 & 0 \end{pmatrix}$$

is stable.

Let

$$\tilde{X} = X + \begin{pmatrix} 0.3 & -0.2 \\ -0.2 & 0.8 \end{pmatrix} \times 10^{-j}$$

be an approximate solution to the DARE (1.2). Take $\alpha = \|F\|_F$, $\beta = \|G\|_F$, $\gamma = \|H\|_F$ in Theorem 2.1, (3.2), and (3.9). Some numerical results on lower and upper bounds for the backward error $\eta(\tilde{X})$ are listed in Table 1 and Table 2, where ψ denotes the relative condition number of the DARE (1.2) defined by (3.11). The cases when the condition (2.50) of Theorem 2.1 is violated are denoted by an asterisk.

From the results listed in Table 1 we see that the conditioning of the DARE (1.2) of this example deteriorates as m increases, and the backward error of $\tilde{X}$ increases as the conditioning of the DARE deteriorates. Note that the symmetric p.s.d. solution

$$X = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

to the DARE (1.2) of this example is independent of m.

From the results listed in Table 2 we see that the backward error of $\tilde{X}$ increases as j decreases (i.e., as the error $\|\tilde{X} - X\|_F$ increases).

TABLE 1

$j = 13$, $\|\tilde{X} - X\|_F = 9.00\text{E} - 14$, $\|X\|_F = 1\ \forall m$

m	ψ	$l(\rho)/\sqrt{3}$	$u(\rho)$	$\rho/\sqrt{3}$	ρ	η_R	$2.12\eta_R$
0	1.62E + 00	3.74E − 14	6.48E − 14	3.74E − 14	6.48E − 14	6.48E − 14	1.37E − 13
1	1.00E + 02	1.69E − 12	2.92E − 12	1.69E − 12	2.92E − 12	2.92E − 12	6.19E − 12
2	1.00E + 04	1.73E − 10	3.00E − 10	1.73E − 10	3.00E − 10	3.00E − 10	6.36E − 10
3	1.00E + 06	1.68E − 08	3.10E − 08	1.73E − 08	3.00E − 08	3.00E − 08	6.36E − 08
4	1.00E + 08	*	*	1.73E − 06	3.00E − 06	3.00E − 06	6.36E − 06

TABLE 2

$m = 2$, $\psi = 1.00\text{E} + 04$, $\|X\|_F = 1$

j	$\|\tilde{X} - X\|_F$	$l(\rho)/\sqrt{3}$	$u(\rho)$	$\rho/\sqrt{3}$	ρ	η_R	$2.12\eta_R$
13	9.00E − 14	1.73E − 10	3.00E − 10	1.73E − 10	3.00E − 10	3.00E − 10	6.36E − 10
11	9.00E − 12	1.73E − 08	3.00E − 08	1.73E − 08	3.00E − 08	3.00E − 08	6.36E − 08
9	9.00E − 10	1.68E − 06	3.10E − 06	1.73E − 06	3.00E − 06	3.00E − 06	6.36E − 06
8	9.00E − 09	*	*	1.73E − 05	3.00E − 05	3.00E − 05	6.36E − 05

EXAMPLE 4.2 [8]. Consider the DARE (1.2) with

$$F = VF_0V, \qquad G = VG_0V, \qquad H = VH_0V,$$

where

$$F_0 = \operatorname{diag}(0, 10^{-m}, 1), \qquad G_0 = \operatorname{diag}(10^{-m}, 10^{-m}, 10^{-m}),$$

$$H_0 = \operatorname{diag}(10^{m}, 1, 10^{-m}),$$

and

$$V = I - \frac{2vv^T}{3}, \qquad v = (1, 1, 1)^T.$$

The unique symmetric p.s.d. solution X to the DARE (1.2) is given by $X = VX_0V$, where $X_0 = \operatorname{diag}(x_1, x_2, x_3)$ with

$$x_i = \frac{f_i^2 + h_i g_i - 1 + \left[\left(f_i^2 + h_i g_i - 1\right)^2 + 4h_i g_i\right]^{1/2}}{2g_i},$$

and h_i, f_i, and g_i are the corresponding diagonal elements of H_0, F_0, and G_0.

Let

$$\tilde{X} = X + \begin{pmatrix} 0.5 & -0.1 & 0.2 \\ -0.1 & 0.3 & 0.6 \\ 0.2 & 0.6 & -0.4 \end{pmatrix} \times 10^{-j}$$

be an approximate solution to the DARE (1.2). Take $\alpha = \|F\|_F$, $\beta = \|G\|_F$, $\gamma = \|H\|_F$ in Theorem 2.1, (3.2), and (3.9). Some numerical results on lower and upper bounds for the backward error $\eta(\tilde{X})$ are listed in Table 3 and Table 4, where ψ denotes the relative condition number of the DARE (1.2) defined by (3.11). The cases when the condition (2.50) of Theorem 2.1 is violated are denoted by an asterisk.

From the results listed in Table 3 we see that the conditioning of the DARE (1.2) of this example deteriorates and the magnitude of the symmetric p.s.d. solution X increases as m increases.

From the results listed in Tables 1–4 we see that the linear estimates (4.2) [i.e., (2.54) and (3.4)] are relatively sharp, while the nonlinear estimates

TABLE 3

$j = 9, \|\tilde{X} - X\|_F = 1.15\text{E} - 09$

m	ψ	$\|X\|_F$	$l(\rho)/\sqrt{3}$	$u(\rho)$	$\rho/\sqrt{3}$	ρ	η_R	$2.12\eta_R$
0	1.20E + 00	2.50E + 00	2.83E − 10	4.89E − 10	2.83E − 10	4.89E − 10	4.04E − 10	8.57E − 10
1	5.24E + 00	1.01E + 01	6.23E − 11	1.08E − 10	6.23E − 11	1.08E − 10	1.01E − 10	2.15E − 10
2	4.71E + 01	1.00E + 02	6.55E − 12	1.13E − 11	6.55E − 12	1.13E − 11	1.13E − 11	2.39E − 11
3	4.66E + 02	1.00E + 03	6.58E − 13	1.14E − 12	6.58E − 13	1.14E − 12	1.14E − 12	2.42E − 12
4	4.66E + 03	1.00E + 04	6.59E − 14	1.14E − 13	6.59E − 14	1.14E − 13	1.14E − 13	2.42E − 13
5	4.66E + 04	1.00E + 05	6.54E − 15	1.13E − 14	6.54E − 15	1.13E − 14	1.13E − 14	2.40E − 14

TABLE 4

$m = 0, \psi = 1.20\text{E} + 00, \|X\|_F = 2.50\text{E} + 00$

j	$\|\tilde{X} - X\|_F$	$l(\rho)/\sqrt{3}$	$u(\rho)$	$\rho/\sqrt{3}$	ρ	η_R	$2.12\eta_R$
9	1.15E − 09	2.83E − 10	4.89E − 10	2.83E − 10	4.89E − 10	4.04E − 10	8.57E − 10
7	1.15E − 07	2.83E − 08	4.89E − 08	2.83E − 08	4.89E − 08	4.04 − 08	8.57E − 08
5	1.15E − 05	2.83E − 05	4.89E − 06	2.83E − 06	4.89E − 06	4.04E − 06	8.57E − 06
3	1.15E − 03	2.82E − 04	4.90E − 04	2.83E − 04	4.89E − 04	4.04E − 04	8.57E − 04
1	1.15E − 01	2.63E − 02	5.24E − 02	2.83E − 02	4.90E − 02	4.05E − 02	8.59E − 02
0	1.15E + 00	*	*	2.90E − 01	5.02E − 01	4.16E − 01	8.83E − 01

(4.1) [i.e., (2.51)] even do not exist in some cases. However, it is worth pointing out that the nonlinear estimates (4.1) guarantee the existence of the solution to the optimization problem (1.3) with the constraints (1.4)–(1.6), while the linear estimates (4.2) would formally give approximate bounds which might not correspond to any solution to (1.3)–(1.6).

I would like to thank the referees for helpful comments and suggestions.

REFERENCES

1 B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, N.J., 1979.
2 W. F. Arnold, III, and A. J. Laub, Generalized eigenproblem algorithms and software for algebraic Riccati equations, *Proc. IEEE* 72:1746–1754 (1984).
3 A. R. Ghavimi and A. J. Laub, Backward error, sensitivity, and refinement of computed solutions of algebraic Riccati equations, *Numer. Linear Algebra Appl.* 2:29–49 (1995).
4 A. Graham, *Kronecker Products and Matrix Calculus with Applications*, Wiley, New York, 1981.
5 T. Gudmundsson, C. Kenney, and A. J. Laub, Scaling of the discrete-time algebraic Riccati equation to enhance stability of the Schur solution method, *IEEE Trans. Automat. Control* 37:513–518 (1992).
6 N. J. Higham, Perturbation theory and backward error for $AX - XB = C$, *BIT*, 33:124–136 (1993).
7 D. J. Higham and N. J. Higham, Backward error and condition of structured linear systems, *SIAM J. Matrix Anal. Appl.* 13:162–175 (1992).
8 M. M. Konstantinov, P. Hr. Petkov, and N. D. Christov, Perturbation analysis of the discrete Riccati equation, *Kybernetika* 29:18–29 (1993).
9 B. Kågström, A perturbation analysis of the generalized Sylvester equation $(AR - LB, DR - LE) = (C, F)$, *SIAM J. Matrix Anal. Appl.* 15:1045–1060 (1994).
10 A. J. Laub, A Schur method for solving algebraic Riccati equations, *IEEE Trans. Automat. Control* AC-24:913–921 (1979).
11 V. Mehrmann, A symplectic orthogonal method for single input or single output discrete time optimal quadratic control problems, *SIAM J. Matrix Anal. Appl.* 9:221–247 (1988).
12 J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic, New York, 1970.
13 T. Pappas, A. J. Laub, and N. R. Sandell, Jr., On the numerical solution of the discrete-time algebraic Riccati equation, *IEEE Trans. Automat. Control* AC-25:631–641 (1980).
14 J. L. Rigal and J. Gaches, On the computability of a given solution with the data of a linear system, *J. Assoc. Comput. Mach.* 14:90–101 (1967).

15 J.-G. Sun, Residual Bounds of Approximate Solutions of the Discrete-Time Algebraic Riccati Equation, Report UMINF 95.17, ISSN-0348-0542, Dept. of Computing Science, Umeå Univ., 1995.

16 J.-G. Sun, Perturbation Theory for the Discrete-Time Algebraic Riccati Equation, Report UMINF 95.20, ISSN-0348-0542, Dept. of Computing Science, Umeå Univ., 1995.

Received 2 October 1995; final manuscript accepted 13 May 1996